ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2025.0177

Attention mechanism-based multi-domain feature fusion approach for active sonar target recognition

Tongjing SUN[1,2*], Haoran XU[1,2], Shishuo REN[1,2], Denghui ZHANG[1,2]

[1]*School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China*
[2]*Key Laboratory of Communication Information Transmission and Fusion Technology, Hangzhou Dianzi University, Hangzhou 310018, China*

Abstract: Due to the complex and changeable marine environment, the active sonar target recognition problem has always been difficult in the field of underwater acoustics. Deep learning-based fusion recognition technology provides an effective way to solve this problem, but relying on simple concatenation strategies to fuse multi-domain features can cause information redundancy, and it isn't easy to effectively mine correlation information between domains. Therefore, this paper proposes an attention mechanism-based multi-domain feature fusion recognition method for active sonar targets. In the method, by preprocessing active sonar echo signals and constructing a multi-domain feature extraction and fusion network, this network uses a 1-DCNN-LSTM series network and a 2-DCNN network with channel attention introduced respectively to extract deep features from different domains. Subsequently, combining feature concatenation and constructing multi-domain cross-attention, intra-domain and cross-domain feature fusion are performed, which can effectively eliminate redundant information and promote inter-domain information interaction while maximizing the retention of target features. Experimental results show that compared with single-domain methods, the network using an attention mechanism for multi-domain feature fusion strengthens cross-domain information interaction and significantly improves feature representation capability. Compared with other methods, the recognition method in this paper also has obvious advantages in performance and maintains stable generalization ability in scenarios with a low signal-clutter ratio.

Key words: Acoustic target recognition; Neural network; Attention mechanism; Multi-domain feature fusion

Unedited

1 Introduction

With the development of marine resource exploitation and national defense technology, the demand for underwater acoustic target recognition using sonar continues to grow (Li, 2012). Active sonar can achieve high-precision localization and recognition of targets in various underwater detection scenarios, holding significant importance in applications such as underwater search and rescue, fishing, seabed environment detection, and frogman/submarine detection (Huang et al., 2019). However, the complex and variable marine environment poses numerous challenges for underwater acoustic target recognition (Domingos et al., 2022). First, during sound wave propagation, multipath effect occur due to reflections from the sea surface and seabed (Fang et al., 2019), while moving targets such as submarines, torpedoes, and underwater vehicles generate Doppler effects (Khan et al., 2024), leading to distortion of echo signals. Second, various types of clutter interference and noise exist in underwater environments, which may cover or confuse target signals, increasing recognition difficulty. Additionally, the complexity and difficulty of underwater exploration result in a scarcity of samples, constraining the feature extraction capability of models and making it challenging to fully extract target feature information. Facing these problems, researchers both domestically and internationally have been continuously committed to research in this field.

Early underwater acoustic target recognition relied on manual listening to analyze the characteristics of acoustic signals (Arrabito et al., 2005), but this approach is susceptible to environmental and personnel factors, greatly limiting its recognition rate. With the development of computer science and the proposal and validation of signal

✉Tongjing SUN, stj@hdu.edu.cn
Haoran XU, https://orcid.org/0009-0007-2234-9683
Shishuo REN, https://orcid.org/0009-0006-3650-048X
Denghui ZHANG, https://orcid.org/0009-0006-3310-3358
CLC number:

Received: Dec. 13, 2025; Revision accepted: Jan. 13, 2026; Crosschecked: Jan. 23, 2026

© The Authors 2026. Published by Zhejiang University Press Co., Ltd. This is an open access article distributed under the terms of the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

processing techniques, the advancement of underwater target recognition technology has been promoted. Researchers adopted machine learning methods to classify targets based on feature extraction. Shin et al. (1997) observed and analyzed the relationship between target echo signals and target shape/internal structure, extracting multiple features including time-domain waveforms, frequency spectra, time-frequency spectra, and auditory perceptual characteristics. By fusing the feature information and combining various machine learning algorithms such as neural networks, they achieved correct classification of target echoes and clutter. Young et al. (2007) extracted multiple auditory perceptual features from target echo signals, performed feature screening using principal component analysis, and successfully distinguished echoes from offshore oil drilling platforms and cruise ships from various surrounding clutter by employing a Bayesian classifier model. However, traditional machine learning algorithms based on manual features—such as Support Vector Machine (SVM) and K-Nearest Neighbors (KNN)—require the reliance on expert experience or prior knowledge. Their capability to process large-scale, high-dimensional data is limited, making it difficult to fully mine deep-level information in the data, resulting in unsatisfactory target classification accuracy rates in complex underwater environments (Hu et al., 2018).

With the rapid development of artificial intelligence technology, deep learning can automatically learn deep-level features from massive data (Liu et al., 2020), eliminating the need for cumbersome manual feature design. This provides an effective approach for underwater acoustic target recognition and has garnered widespread attention. Researchers have mainly conducted studies from two aspects: feature representation (Shadlou et al., 2019; Han et al., 2022; Yang et al., 2016) and fusion strategies (Zhang et al., 2018; Hong et al., 2021; Hu et al., 2021). At the feature extraction level, early researchers focused on using neural networks for deep feature mining in a single domain. Lee et al. (2020) employed Power Normalized Cepstral Coefficients to represent time-domain features, combined with Convolutional Neural Network (CNN) to accomplish target classification. Choo et al. (2022) performed Short-Time Fourier Transform (STFT) time-frequency processing on a small number of real echo signals and achieved a high echo detection rate using a designed shallow neural network for target detection.

Given that echo signals are complex time-varying signals with strong inter-frame correlations, some researchers have attempted to use memory-capable neural networks such as RNN and LSTM to process such signals. For example, Kamal et al. (2021) proposed a model combining 1DCNN and LSTM, specifically introducing a selective attention layer to focus on key features. In experiments on collected hydroacoustic data, this method achieved 95.2% recognition accuracy, demonstrating the advantage of incorporating attention mechanisms in feature extraction. However, the representational capability of single-domain features remains limited by signal dimensionality, making it difficult to distinguish targets in complex environments. Pan et al. (2024) constructed a multimodal deep learning model for ship-radiated noise recognition, feeding one-dimensional time series and two-dimensional time-frequency images into a joint network to enhance recognition accuracy through cross-modal feature interaction. Wang et al. (2024) proposed an active target recognition method based on multi-domain active target echo images and attention networks, utilizing a deep neural network with attention mechanisms to learn multi-domain joint feature representations, significantly improving target recognition accuracy compared to single-domain inputs.

Although many valuable achievements have been obtained, the combination of attention mechanisms and multi-domain fusion has shown potential in the field of underwater acoustic target recognition. However, most existing methods adopt dual-domain interaction or single attention fusion approaches, failing to sufficiently explore the complex dependencies between cross-domain features, which leads to issues of feature redundancy and insufficient utilization of complementarity. Notably, attention mechanism-based multi-domain feature fusion has formed a mature application paradigm in fields such as computer vision and radar target recognition. He et al. (2023) proposed the EnVLF framework for remote sensing cross-modal text-image retrieval, which integrates a vision transformer module and a multi-modal encoder with cross-modal attention to align visual-text semantics and improve retrieval performance. Li et al. (2024) developed a multi-modal sentiment analysis model for image-text data: it uses DenseNet121 and CBAM to extract image emotional features and uses ALBert and BiLSTM to extract text features, then adopts cross-attention for feature correlation modeling, enhancing classification accuracy. Gao et al. (2025) designed for multi-source remote sensing classification, extended Mamba to adapt to heterogeneous inputs (hyperspectral image and SAR data), and enhanced cross-modal interaction via attention-like modeling, boosting fusion and classification effects. While these attention-based multi-domain/multi-modal fusion methods have achieved success in their fields, there remains ample room for exploration of such methods in active sonar target recognition. Building upon the aforementioned research, this paper proposes an attention mechanism-based multi-domain feature fusion recognition method for active sonar targets. By constructing a multi-domain cross-attention mechanism, it achieves cross-domain information interaction among time-domain envelope features, time-azimuth features, and time-frequency features, fully mining the multi-domain information of active sonar signals, thereby

Unedited

improving the accuracy of underwater acoustic target recognition and classification.

2 Overall framework of the method

The overall recognition flow of the attention mechanism-based multi-domain fusion recognition method for active sonar targets is shown in Fig. 1. This method consists of three modules: data preprocessing, multi-domain feature extraction and fusion, and classification recognition. First, data preprocessing is performed on the active sonar echo signals to analyze different domains of the signals and provide inputs for subsequent feature extraction networks. Subsequently, through a multi-domain feature extraction and fusion network incorporating an attention mechanism, deep feature extraction and fusion of the target's multi-domain features are conducted to excavate inter-domain correlations and strengthen key information. Finally, the classification and recognition module performs classification and recognition on the fused multi-domain joint features.

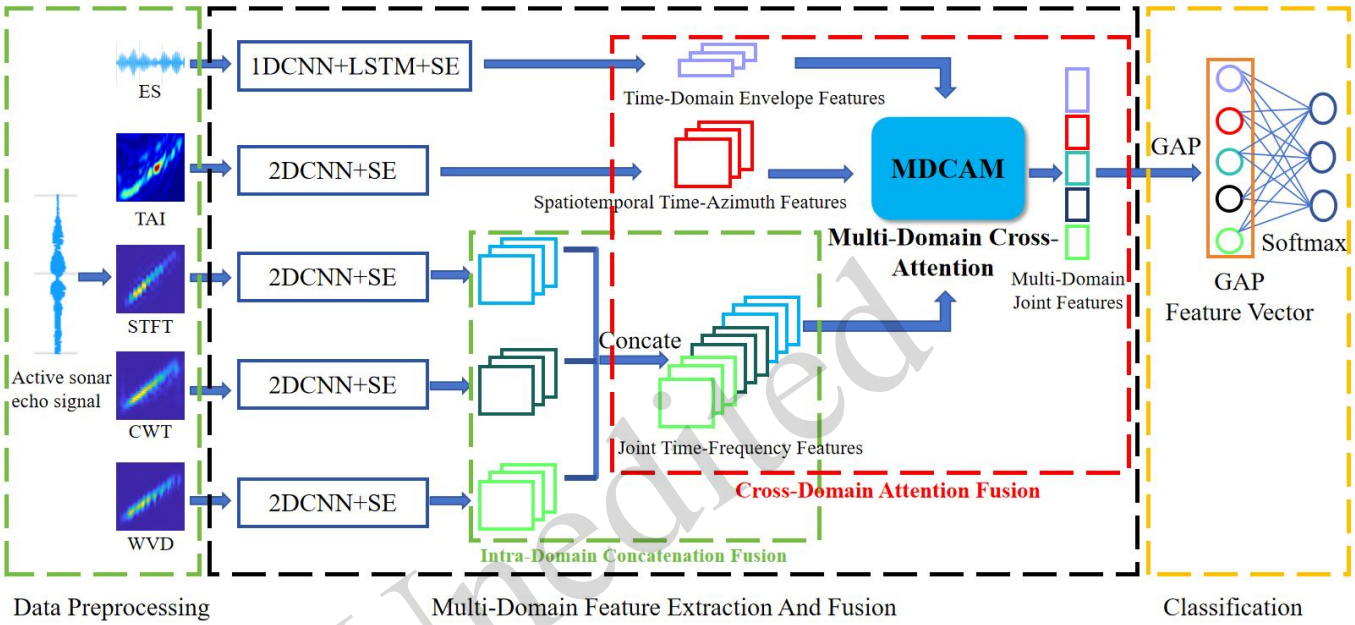


Fig. 1 Overall recognition workflow of the proposed method

In the data preprocessing module, the active sonar echo signals undergo time-domain, space-time domain, and time-frequency domain analysis, converting them into time-domain Envelope Sequences (ESs), space-time domain Time Azimuth Images (TAIs) and time-frequency domain images. For time-domain analysis, the modulus of the analytic signal is calculated via the Hilbert transform to extract the time-domain envelope sequence, reducing the influence of high-frequency carrier periodic oscillations while preserving low-frequency modulation information reflecting the slow-varying trends of target echo energy. In space-time domain analysis, leveraging the advantage of prior knowledge about transmitted signals in active sonar echoes, matched filtering and space-time conversion are applied to generate two-dimensional TAIs, which effectively characterizes the dynamic spatial-azimuth properties of targets. Because neither the time domain nor the frequency domain alone can capture the time-varying frequency characteristics of signals, and the instantaneous frequency evolution of target echoes constitutes key discriminative information, three time-frequency transformation methods—Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), and Wigner-Ville Distribution (WVD)—are jointly employed for processing target echo signals in time-frequency domain analysis.

In the multi-domain feature extraction and fusion module, firstly, 1-DCNN-LSTM and 2-DCNN networks are designed to extract features from different domains of signals, with the Squeeze and Excitation (SE) (Hu et al., 2018) channel attention mechanism incorporated into each to perform adaptive weighting on the channel dimension of features from different domains. Its core idea is squeezing global feature statistical information and prompting the adaptive adjustment of channel weights, thereby strengthening key features and suppressing redundant information. Secondly, to overcome the limitations of traditional static fusion through simple concatenation and fully utilize complementary information from different domains of echo signals, feature concatenation is initially applied for intra-domain feature fusion, followed by the construction of a Multi-Domain Cross-Attention Mechanism (MDCAM). This mechanism performs cross-attention computation across multi-domain features, dynamically learning importance weights for features from different domains, thereby maximizing target feature retention while effectively eliminating redundant

Unedited

information and promoting inter-domain information interaction.

In the classification and recognition module, a lightweight classification architecture combining Global Average Pooling (GAP) and Softmax is adopted for accurate and efficient classification of the fused multi-domain joint features. Unlike traditional Fully Connected (FC) layers, where excessive trainable parameters easily cause model overfitting, GAP aggregates feature maps through global mean pooling, mapping high-dimensional features to low-dimensional vectors without introducing additional trainable parameters, significantly enhancing model generalization capability. First, GAP operation is performed on the fused multi-domain joint features to generate feature vectors containing global information. Subsequently, these vectors are fed into the Softmax layer, where class probability distributions are computed via the normalized exponential function to achieve target classification. This design leverages the parameter-free nature and global information integration capability of GAP, reducing model complexity while preserving feature discriminability, effectively preventing overfitting, and ensuring that underwater target classification results possess both accuracy and interpretability.

The multi-domain feature extraction and fusion module constitutes the key research focus of this paper, and will be explained in detail below.

2.1 Multi-domain feature extraction based on channel attention

Aiming at the different representation forms of active sonar echo signals, differentiated deep network architectures are constructed to achieve hierarchical feature mining, and the channel attention mechanism is employed to enhance focus on important features. Specifically, a 1DCNN-LSTM serial network is adopted to extract time-domain envelope features of echo signals, while 2DCNN networks are utilized to extract space-time domain time-azimuth features and time-frequency domain time-frequency features, respectively. The SE channel attention mechanism (Hu et al., 2018) is embedded into each network, adaptively adjusting channel weights through global pooling and nonlinear mapping to strengthen target-relevant features.

Among them, the SE mechanism effectively addresses the information redundancy problem caused by convolutional layers learning channel features through kernels. By adaptively assigning weights to each channel of the feature map, it enables the network to focus on features more critical for target recognition, thereby enhancing channel-wise feature representation and suppressing interference from redundant information. Its structure is shown in Fig. 2, primarily composed of three parts: Squeeze, Excitation, and Re weighting. Assuming input feature X, where H is the feature map height, W is the width, and C is the number of channels. First, the compression operation of the SE module, namely Global Average Pooling (GAP) applied to the input feature map, compresses spatial dimensions into channel descriptors. Subsequently, the excitation step uses two fully connected layers (FC) to perform a nonlinear transformation on the channel descriptors, learning dependencies between channels, with ReLU activation in the middle, and finally Sigmoid to generate weights. The reweighting step multiplies the channel weights with the original feature map channel by channel, enhancing key channel features and suppressing redundant channels, obtaining the weighted feature map.

2.1.1 1DCNN-LSTM-SE time-domain feature extraction network

Traditional methods often use images transformed by time-frequency analysis as network model inputs, leading to loss of original signal time-domain information. This paper takes the envelope sequence processed from active sonar echo signals through time-domain analysis as input. This envelope sequence focuses on the slow-varying trends of signal amplitude, effectively removing high-frequency carrier interference, and can completely preserve the time-domain dynamic information of target echoes. Aiming at the one-dimensional time series characteristics of the envelope sequence, a 1DCNN-LSTM-SE time-domain feature extraction network is constructed, with its structure shown in Fig. 3.

In the time-domain feature extraction network, the 1DCNN part consists of alternating convolutional layers and max-pooling layers. The convolutional layers contain 8, 16, and 32 convolution kernels with a size of 3, respectively, extracting local temporal features in the envelope sequence through sliding window operations. Its translation invariance effectively preserves the time-domain continuity of envelope features, avoiding feature extraction deviations caused by data shifts. The max-pooling layers use a pooling window of size 3 to reduce the dimensionality of the convolutional layers' outputs and lower computational complexity while preventing overfitting. Batch Normalization (BN) layers and ReLU activation functions are added after the convolutional and pooling layers to prevent overfitting and improve model stability. This modular stacking design draws on the structural idea of ResNet, enabling multi-scale feature extraction through stacked modules, thus efficiently learning time-domain features under limited resources while keeping the network lightweight. Subsequently, the SE channel attention mechanism enhances the features output by the 1DCNN, dynamically strengthening target-relevant time-domain features through global pooling, channel weight learning, and feature re-weighting. Finally, LSTM is used to further enhance the feature extraction capability for long-term temporal dependencies in the envelope sequence. It contains

Unedited

64 memory units, selectively retaining or forgetting envelope feature information through gating mechanisms, thereby effectively capturing long-term dependencies in the envelope sequence. Its output feature vector serves as the complete time-domain envelope feature representation for subsequent feature fusion and target recognition tasks.

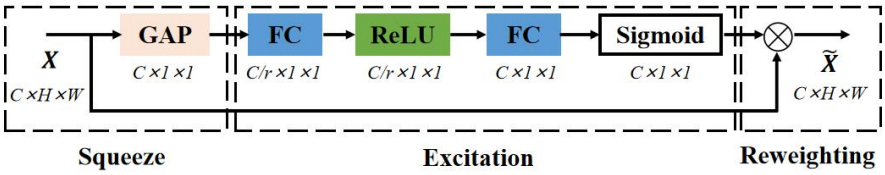


Fig. 2 Structure diagram of the SE channel attention mechanism

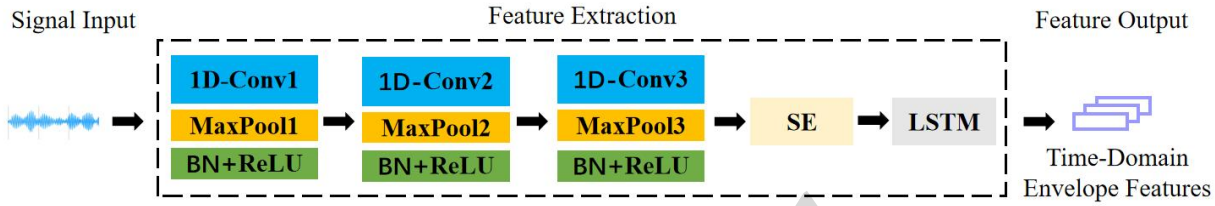


Fig. 3 1DCNN-LSTM-SE time-domain feature extraction network structure diagram

2.1.2 2DCNN-SE space-time-frequency domain feature extraction network

For the two-dimensional space-time images of time-azimuth and time-frequency images generated after space-time domain and time-frequency domain analysis of active sonar echo signals, this paper constructs a 2DCNN-based network structure. To enhance the network's feature extraction capability for space-time-frequency domain image features, 2DCNN is similarly combined with the channel attention SE mechanism.

By integrating the structural advantages of 2DCNN and SE, a 2DCNN-SE deep feature extraction network is ultimately constructed, with its network structure shown in Fig. 4. The 2DCNN part consists of alternating convolutional layers and max-pooling layers. The convolutional layers contain 8, 16, and 32 convolution kernels with a size of 3×3, respectively. The max-pooling layers use a 2×2 pooling window with a stride all set to 2, and Batch Normalization (BN) layers and ReLU activation functions are added after the convolutional and pooling layers. Similarly, this hierarchical extraction of features at different scales via multiple modules can efficiently enhance feature representation capability and achieve lightweight deep feature learning. The SE part, while maintaining the deep feature extraction capability of 2DCNN, dynamically adjusts the weights of each channel through internal GAP and FC layers, enhancing key feature extraction capability and suppressing noise and redundant information.

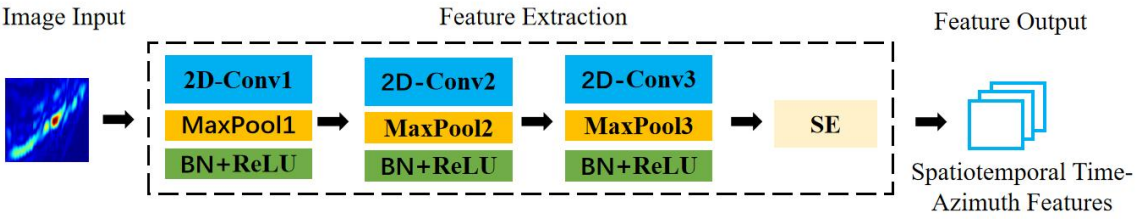


Fig. 4 Structure diagram of 2DCNN-SE spatio-temporal-frequency domain feature extraction network

2.2 Multi-domain feature fusion based on cross-attention

To fully utilize the time-domain envelope features of target echo signals, the time-azimuth features in the spatiotemporal domain, and the time-frequency features from different time-frequency analyses, we fuse both intra-domain and cross-domain features to achieve a more comprehensive representation and improve recognition performance. While multi-domain features provide diverse target characterizations, they may also introduce redundancy and increase computational complexity. Therefore, the key challenge in feature fusion is to maximize complementary information while reducing redundancy under limited hydroacoustic data conditions. In this study, we design intra-domain and cross-domain fusion strategies within the multi-domain feature fusion module. For intra-domain fusion, time-frequency features extracted from different time-frequency analyses are combined via feature concatenation to enhance single-domain representations and

Unedited

form a joint time-frequency feature. For cross-domain fusion, we employ an attention mechanism to adaptively weight and fuse features, learning inter-domain dependencies while suppressing redundancy and enabling complementary information integration. The resulting multi-domain joint feature provides richer discriminative information for target classification, significantly improving recognition performance in complex environments.

2.2.1 Intra-domain feature concatenation fusion

Prior to cross-domain fusion using multi-domain cross-attention, the time-frequency features extracted from STFT, CWT, and WVD time-frequency analyses via the 2DCNN-SE feature extraction network are fused through feature concatenation to form a joint time-frequency representation, maximizing the retention of time-frequency characteristics of the target. Specifically, feature concatenation is performed along the channel dimension using a concatenation fusion layer, resulting in a unified time-frequency feature representation. Assuming the feature maps extracted from the three time-frequency analysis methods are denoted $F_i \in R^{C_i \times H \times W}$ (where i=1,2,3), with H, W representing spatial dimensions and C_i representing the channel number, the concatenated output feature is

$$F_{concat} = F_1 \oplus F_2 \oplus F_3 \in R^{(C_1+C_2+C_3)\times H \times W} \quad (1)$$

where $\oplus$ denotes the concatenation operation along the channel dimension, merging features from different time-frequency analysis methods. This process preserves the feature information from various time-frequency representations, providing a foundation for subsequent cross-domain fusion.

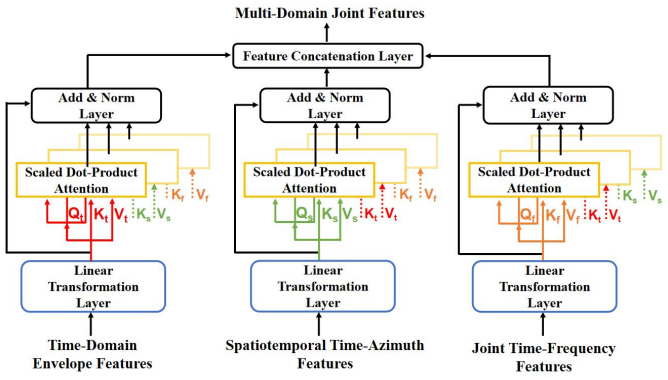
2.2.2 Cross-domain attention mechanism fusion

To address the challenges of distribution, scale, and type discrepancies across different domains, as well as the redundancy and computational complexity introduced by multi-feature fusion, traditional fusion methods often fail to fully exploit inter-domain dependencies, leading to insufficient information utilization. The attention mechanism, due to its ability to capture feature dependencies, offers an effective solution for multi-domain feature fusion. Therefore, this study proposes a Multi-Domain Cross-Attention Module (MDCAM), an extension of the Transformer's multi-head attention mechanism (Vaswani et al., 2017). By combining self-attention and cross-domain attention, MDCAM enables deep interaction and adaptive fusion of time-domain, spatiotemporal-domain, and time-frequency-domain features. The module consists of linear transformation layers, scaled dot-product attention, normalization layers, and feature concatenation layers, as illustrated in Fig 5. Through MDCAM, the model learns nonlinear inter-domain dependencies, suppresses redundant information, enhances complementary features, and achieves efficient multi-domain feature fusion.

Specifically, the temporal envelope features, spatiotemporal time-azimuth features, and joint time-frequency features are input into their respective linear transformation layers to generate corresponding Query (Q), Key (K), and Value (V) matrices, as shown in the temporal features (Q_t,K_t,V_t), spatiotemporal features (Q_s,K_s,V_s), and time-frequency features (Q_f,K_f,V_f), in the figure. Subsequently, the computation proceeds to multi-head attention, which consists of two parts: self-attention computation and cross-domain attention computation. In self-attention computation, features within each domain undergo scaled dot-product attention calculation to capture intra-domain feature dependencies and enhance the expression of key information. In cross-domain attention computation, the Q from one domain interacts with the K and V from other domains to explore potential inter-domain correlations and achieve cross-domain information complementarity. Taking the temporal envelope feature processing path as an example, assuming the input temporal feature X has dimension d, after passing through the linear transformation layer, the generated Q_t and the K_s, V_s from the spatiotemporal feature path are input into the scaled dot-product attention for cross-domain attention computation. The computation process can be expressed as

$$Attention(Q_t, K_s, V_s) = V_s * Softmax\left(\frac{Q_t * K_s^T}{\sqrt{d_{K_s}}}\right) \quad (2)$$

Here, Q_t, K_s, and V_s are obtained through linear transformations using three different weight matrices W_q, W_k, and W_v, respectively, e.g., $Q_t = XW_q$. By multiplying Q_t with the K_s matrix, a correlation score matrix is obtained, which serves as a similarity measure between the temporal and spatiotemporal domains. This matrix is then normalized via Softmax to produce attention weights, which are used to weight V_s, enabling the temporal features to incorporate information from the spatiotemporal features. After multi-head attention computation, the outputs of the multi-head attention are added to the input features and normalized. Finally, the features are integrated through a concatenation layer to produce the joint multi-domain feature output.



Unedited

Fig. 5 Schematic diagram of feature fusion in the multi-domain cross-attention module

The cross-domain feature fusion approach based on multi-domain cross-attention not only enhances inter-domain interaction, allowing features from different domains to guide and complement each other, but also enables the model to adaptively filter irrelevant information and focus on complementary features through the synergistic effect of self-attention and cross-domain attention. Moreover, compared to traditional multi-branch attention structures, MDCAM significantly reduces the number of parameters by sharing the multi-head attention module while maintaining performance and improving computational efficiency.

3 Method validation

3.1 Experimental description and data analysis

The active sonar echo signal experiment was conducted at the Mogan Lake Underwater Acoustic Test Site. The transmitted signal employed a Linear Frequency Modulated (LFM) waveform with a pulse width of 2 ms and a frequency range of 40-80 kHz, sampled at 1 MHz. Fig. 6 illustrates the experimental target placement configuration. The transducer, hydrophone, and target were positioned at the same depth, with the hydrophone placed between the transducer and the target rotation platform. The target was slowly rotated via the platform. The sonar signals were emitted by the transducer, propagated toward the target, and reflected echoes were received by the hydrophone.

The experiment used four types of submarine-like cylindrical models as targets: submarine model (Target A), cylinder with hemispherical head and conical tail (Target B), cylinder with both hemispherical head and tail (Target C), and standard cylindrical model (Target D). All four targets had similar materials and shapes, making their features difficult to distinguish in the interference-rich lake test data. The signal waveforms are shown in Fig 7, each with 2400 sampling points. A total of 600-900 signal samples were collected for each target category. After data preprocessing, time-domain ES, space-time domain TAI, and time-frequency images were generated, labeled with category tags to form the sample dataset, which was then divided into training and test sets in a 7:3 ratio.

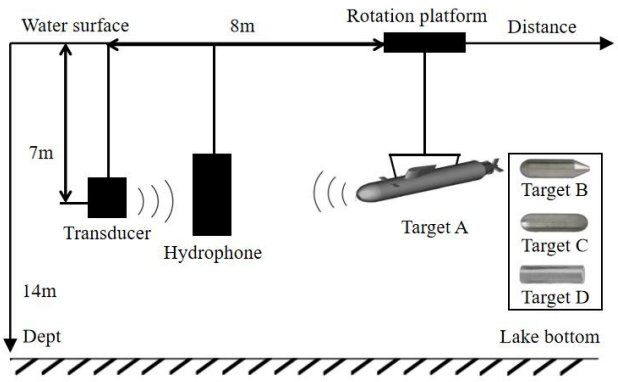
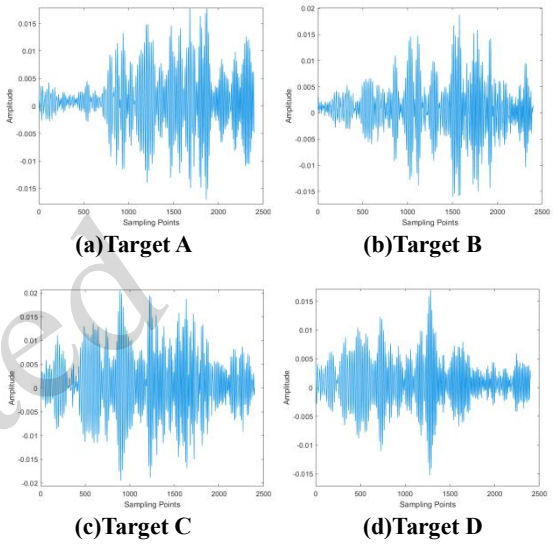


Fig. 6 Schematic diagram of target placement



(a)Target A **(b)Target B**

(c)Target C **(d)Target D**

Fig. 7 Comparison of echo signals from four target types

3.2 Data preprocessing

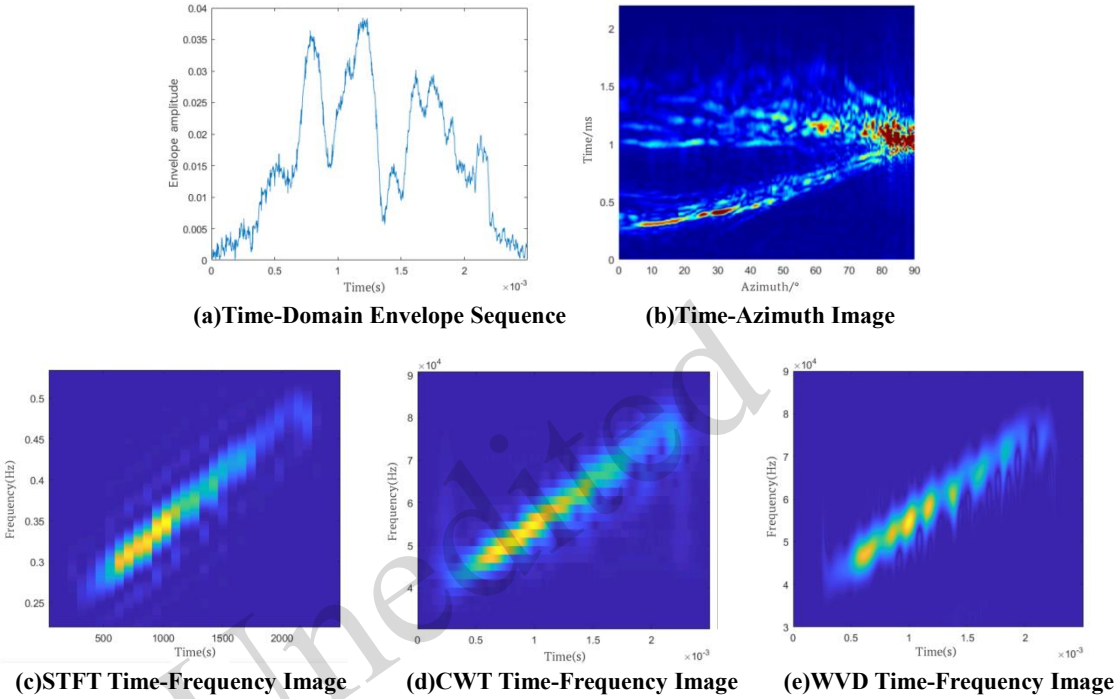
The active sonar echo signals were analyzed in different domains and converted into time-domain ESs, space-time domain TAIs, and time-frequency images. Fig 8 shows the following: (a) The time-domain envelope sequence has a horizontal axis representing the time domain (containing 2400 sampling points) and a vertical axis representing the envelope amplitude, reflecting low-frequency modulation information of the target echo energy's gradual variation trend, (b) The spatiotemporal time-azimuth image has a horizontal axis representing 200 detection beams covering a 90° azimuth angle and a vertical axis representing the signal time series (covering a 2.2ms time range), effectively characterizing the dynamic spatial azimuth characteristics of rotating targets, (c-e) Note the following for the time-frequency domain images: In the STFT time-frequency image, the horizontal axis represents segmented time frames and the vertical axis represents normalized frequency. In the CWT and WVD time-frequency images, the horizontal axis represents time sampling points and the vertical axis represents actual frequency, respectively presenting the frequency variation

Unedited

patterns over time through different time-frequency transforms.

3.3 Experimental results and analysis

All experiments were conducted on an NVIDIA GeForce RTX 3080 server. Models were trained for 100 epochs with full convergence achieved. The batch size was set to 32, and the initial learning rate was 1e−5, halved at the 50th and 75th epochs to facilitate model fine-tuning and prevent overfitting. Optimization used the Adam optimizer. Evaluation metrics included Accuracy(Acc), F1-score(F1), and Recall(R), supplemented by feature visualizations to comprehensively assess network performance on the test set. To validate the effectiveness of the proposed attention-based multi-domain feature fusion method for active sonar target recognition, ablation studies were performed on the attention mechanisms in multi-domain feature extraction and multi-modal feature fusion modules, along with comparative experiments against other network models.



(a)Time-Domain Envelope Sequence (b)Time-Azimuth Image

(c)STFT Time-Frequency Image (d)CWT Time-Frequency Image (e)WVD Time-Frequency Image

Fig. 8 Preprocessing results of target echo signal data

Unedited

3.3.1 Performance validation of integrated channel attention

To verify the effectiveness of incorporating the SE channel attention mechanism in multi-domain feature extraction, we first compared the performance between each basic network and the network integrated with the SE channel attention mechanism. We further introduced mainstream attention mechanisms, such as Convolutional Block Attention Module (CBAM) and Efficient Channel Attention (ECA), for horizontal comparison with SE, so as to clarify the adaptability of different attention mechanisms for the task in this study. Experimental results are shown in Table 1 and Table 2 respectively.

Table 1 Experimental results of multi-domain feature extraction

Domain	Input Data	Network Model	Acc/%	F1/%
Time Domain	Envelope Sequence	1DCNN	82.82	73.14
		LSTM	83.35	73.43
		1DCNN-LSTM	84.05	74.01
		1DCNN-LSTM-SE	**85.71**	**75.38**
Spatiotemporal Domain	Time-Azimuth Image	2DCNN	84.56	73.89
		2DCNN-SE	**86.42**	**75.93**
Time-Frequency Domain	STFT Time-Freq Image	2DCNN	84.90	74.37
		2DCNN-SE	**86.70**	**76.20**

	CWT Time-Freq Image	2DCNN	88.43	78.15
		2DCNN-SE	**90.07**	**79.41**
	WVD Time-Freq Image	2DCNN	86.71	76.83
		2DCNN-SE	**88.49**	**78.56**

From the perspective of channel attention integration in Table 1, when using time-domain ES as input to the LSTM network, its classification accuracy and F1-score slightly outperform 1DCNN, demonstrating LSTM's advantage in capturing long-term temporal dependencies. The series combination of 1DCNN and LSTM further improves performance, indicating complementary integration of local features and global temporal characteristics. After introducing the SE module, the model shows significant enhancement, strengthening the representation capability for time-domain envelope features. For space-time domain TAI and time-frequency domain images, the 2DCNN-SE model with SE attention substantially outperforms conventional 2DCNN, confirming that the SE mechanism adaptively weights channel features to suppress redundant information and focus on critical patterns, validating the effectiveness of attention mechanisms in enhancing discriminative power for 2D feature extraction.

Additionally, the optimal performance metrics across domains in Table 1: The CWT time-frequency image achieves the highest accuracy (90.07%) in the time-frequency domain, highlighting its superiority in capturing fine-grained target features. This stems from the continuous wavelet transform's flexible dilation and translation parameters, which adapt to complex time-frequency variations in echoes and extract subtle signal characteristics. In contrast, STFT offers computational efficiency but limited frequency resolution. WVD provides high time-frequency resolution but suffers from cross-term interference. Both STFT and WVD underperform CWT in comprehensive feature extraction. Time-domain ESs contain discriminative information but yield lower accuracy and F1-scores than other domains, indicating weaker separability and necessitating multi-domain fusion for precise recognition.

Unedited

Table 2 Experimental results of different attention mechanisms

Domain	Input Data	Network Model	Acc/%	F1/%	Parameters/M
Time Domain	Envelope Sequence	1DCNN-LSTM-CBAM	85.83	75.42	9.2
		1DCNN-LSTM-ECA	85.12	74.96	7.8
		1DCNN-LSTM-SE	85.71	75.38	8.1
Spatiotemporal Domain	Time-Azimuth Image	2DCNN-CBAM	86.57	76.01	7.4
		2DCNN-ECA	85.85	75.48	6.1
		2DCNN-SE	86.42	75.93	6.3
Time-Frequency Domain	STFT Time-Freq Image	2DCNN-CBAM	86.72	76.25	7.5
		2DCNN-ECA	86.03	75.62	6.2
		2DCNN-SE	86.70	76.20	6.4

Based on the horizontal comparison results of different attention mechanisms in Table 2, CBAM slightly outperforms the SE mechanism in some models. For instance, the classification accuracy reaches 86.72% when using STFT time-frequency images as input. However, as a hybrid attention mechanism that simultaneously considers both channel attention and spatial attention, CBAM increases the model parameter count by approximately 18% compared with SE. This significant rise in computational complexity hinders its application in real-time sonar systems. Although the ECA mechanism reduces the model parameter count by about 3% through simplifying the channel attention computation, it ignores the nonlinear dependencies between channels. Consequently, it insufficiently captures the complex cross-channel correlations of sonar signals, leading to generally lower performance than SE. In contrast, the SE channel attention mechanism achieves a balance between classification accuracy, model performance, and lightweight

design, making it more suitable for the active sonar target recognition scenario.

3.3.2 Performance validation of integrated cross-attention

To comprehensively evaluate the effectiveness of multi-domain feature fusion and the advantages of the proposed MDCAM in fusion, this subsection conducts experiments in two parts: the first part analyzes the complementarity of features from different domains by comparing single-domain, intra-domain, and cross-domain combinations, as well as multi-domain feature fusion; the second part verifies the unique advantages of MDCAM in cross-domain information interaction and redundancy suppression by introducing other mainstream feature fusion strategies for comparison.

Experimental results of multi-domain feature fusion comparison are shown in Table 3. Taking the single-domain CWT time-frequency image as an example, the 2DCNN-SE model achieves 90.07% accuracy. After feature fusion, performance improves substantially: fusion of CWT and ES via Concat increases accuracy to 90.60%, while MDCAM attention fusion reaches 92.07%. Similarly, the maximum accuracy of combining CWT time-frequency images with space-time domain TAIs rises to 93.49%. For multi-domain fusion, the accuracy of integrating time-frequency images, TAIs, and ESs through hybrid Concat and MDCAM reaches as high as 95.13%.

To intuitively demonstrate the performance differences among different domain combinations, the accuracy of various domains is further compared, as shown in Fig. 9. Compared with single-domain approaches, multi-domain feature fusion achieves significantly superior recognition performance, indicating that synergistic effects are realized by integrating feature information from different domains. Notably, the CWT time-frequency image achieves the highest accuracy among all single-domain inputs. However, intra-domain fusion of CWT and WVD time-frequency images does not yield a significant accuracy improvement, reflecting that intra-domain features may contain substantial redundant information with relatively limited complementarity, making it difficult to achieve a substantial boost in recognition performance. On the other hand, cross-domain fusion of CWT time-frequency images with either ESs or TAIs increases the accuracy by more than 2%, demonstrating the potential of cross-domain information complementarity to enhance feature representation. Additionally, after integrating all time-frequency images from the time-frequency domain with other domains for cross-domain fusion, although the computational burden caused by the additional spectra is slightly increased, the improved accuracy validates the necessity of their integration. It can be inferred from the above experimental analysis that whether additional features from other domains or the same domain can improve recognition performance depends on their ability to provide complementary or unique information. Therefore, achieving synergistic effects through integrating information from different domains facilitates the model in learning more comprehensive and abundant target information, thereby significantly enhancing recognition performance and verifying the significant advantages of multi-domain feature fusion in target recognition.

Unedited

Table 3 Experimental results of multi-domain feature fusion

Domain	Input Data	Feature Extraction	Feature Fusion	Acc/%
Single Domain	CWT Time-Freq Image	2DCNN-SE	–	90.07
Combined Domains	CWT Time-Freq Image & Envelope Sequence	1DCNN-LSTM-SE & 2DCNN-SE	Concat	90.60
	CWT Time-Freq Image & Envelope Sequence	1DCNN-LSTM-SE & 2DCNN-SE	MDCAM	92.07
	CWT Time-Freq Image & Time-Azimuth Image	2DCNN-SE	Concat	91.65
	CWT Time-Freq Image & Time-Azimuth Image	2DCNN-SE	MDCAM	93.49
Multi-Domain	CWT Image, Time-Azimuth Image & Envelope Sequence	1DCNN-LSTM-SE & 2DCNN-SE	Concat	92.52
	CWT Image, Time-Azimuth Image & Envelope Sequence	1DCNN-LSTM-SE & 2DCNN-SE	MDCAM	94.76
	Time-Freq Images*, Time-Azimuth Image & Envelope Sequence	1DCNN-LSTM-SE & 2DCNN-SE	Concat & MDCAM	**95.13**

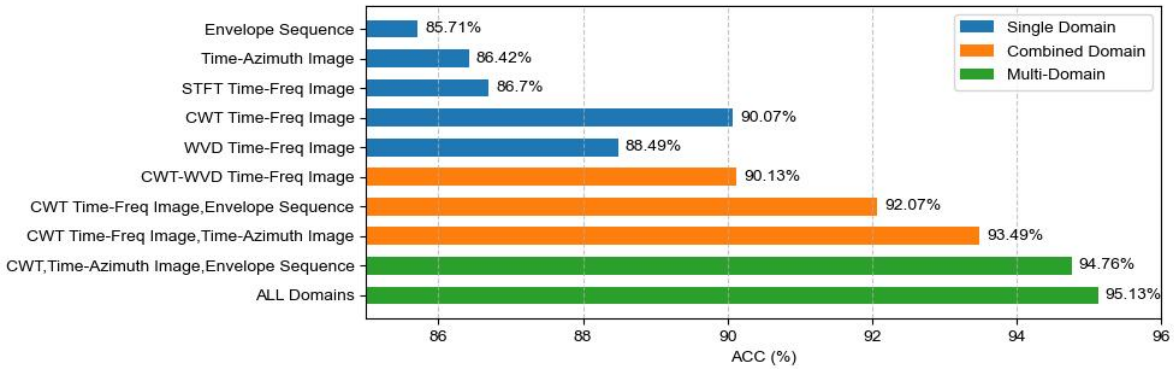
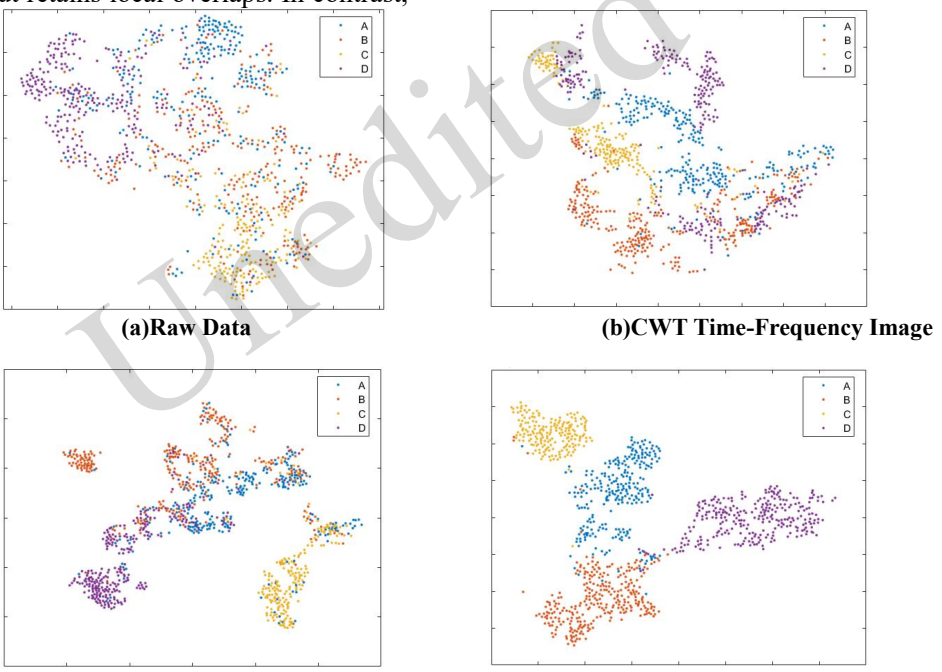


Fig. 9 Comparison of accuracy among single-domain, combined-domain, and multi-domain approaches

To visually validate the classification capability, T-SNE visualizes test set features in Figure 10. In (a), raw data show severe mixing of all four target classes. Using only CWT time-frequency images (b), Class A and B points exhibit extensive overlap, indicating confusion at class boundaries. Introducing cross-domain features (e.g., ESs) improves inter-class separation but retains local overlaps. In contrast, the proposed MDCAM fusion method (d) forms compact clusters with distinct inter-class margins, nearly eliminating A-B overlap. This visually confirms that MDCAM effectively extracts complementary cross-domain features, significantly enhancing discriminative power and demonstrating classification superiority.



(a)Raw Data **(b)CWT Time-Frequency Image**

(c)CWT Time-Frequency Image & Envelope Sequence **(d)Multi-Domain Feature Fusion**

Unedited

Fig. 10 Visualization comparison of output features using different methods

To further verify the unique advantages of MDCAM in cross-domain information interaction and redundancy suppression, other mainstream feature fusion strategies are introduced for comparison in this subsection. The experiments uniformly adopt CWT time-frequency images, ESs, and TAIs as multi-domain inputs, and the comparative experimental results of different feature fusion strategies are shown in Table 4.

Table 4 Comparative experimental results of different feature fusion strategies

Feature Fusion	Acc /%	F1 /%	Parameters /M	FLOPs /G
Concat	92.52	90.35	38.81	1.76
Transformer	93.87	91.72	95.47	3.89
AdaFusion	94.12	91.68	62.76	2.93
MDCAM	**94.76**	**92.12**	58.12	2.37

From the comparative results in Table 4, it can be observed that the traditional Concat fusion method, despite having a slightly lower parameter count, merely stacks

features without exploring inter-domain dependencies, resulting in the weakest feature discriminability. Benefiting from the global modeling capability of the self-attention mechanism, the Transformer fusion strategy achieves a significant accuracy improvement of 93.87%. However, this method requires global attention computation for all features, leading to a high parameter count that makes it difficult to adapt to real-time sonar recognition scenarios. Meanwhile, AdaFusion assigns feature weights through adaptive weighting, outperforming the Transformer fusion. Nevertheless, its weight assignment is only based on statistical information of a single feature dimension, failing to achieve dynamic interaction of cross-domain features, thus still having limitations in complex sonar signal processing. The proposed MDCAM in this paper achieves the optimal performance. This is attributed to its cross-attention computation tailored to the multi-domain characteristics of sonar signals: it can not only adaptively learn the importance weights of features from different domains, strengthen cross-domain information interaction, effectively suppress redundancy, and explore inter-domain complementary features but also avoid the redundant computation of the Transformer's global attention, thereby achieving an optimal balance between performance and efficiency.

3.3.3 Comparative experiments

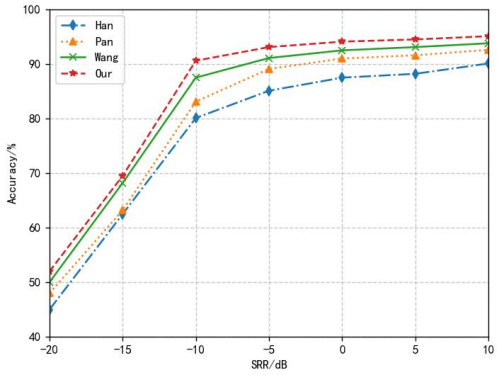
To validate the effectiveness of the proposed attention-based multi-domain feature fusion method for active sonar target recognition, this section compares the proposed method with other models on the test set using recognition accuracy, recall, F1-Score, and parameter count as evaluation metrics. Comparative models include the classical convolutional neural network VGG16 (Simonyan et al., 2014), the self-attention-based network model VIT (Vision Transformer; Dosovitskiy et al., 2020); a method using 1DCNN-LSTM to extract deep temporal features of signals (Han et al., 2022); a method combining 1DCNN-LSTM with 2DCNN and employing SE channel attention for multi-feature fusion (Pan et al., 2024); the method by Wang et al. (2024) that learns multi-domain joint features based on ResNet50 (He et al., 2016) and the Swim Transformer attention mechanism.

Comparative experimental results are shown in Table 5. The proposed method significantly outperforms others in classification accuracy, recall, and F1-Score. Compared with VGG16, accuracy increases by 7.83%, recall improves by 4.05%, F1-Score rises by 4.44%, while model parameters decrease by 48.87M. Versus VIT, accuracy improves by 6.55%, recall increases by 4.79%, and F1-Score rises by 4.16%, achieving superior recognition accuracy with a comparable parameter count. Compared with Han et al.'s time-domain-only method, the proposed multi-domain fusion (incorporating time-domain envelope features, spatiotemporal time-azimuth features, and joint time-frequency features) enhances accuracy by 4.64% despite slightly higher parameters. Relative to Pan et al.'s SE channel attention method and Wang et al.'s Transformer attention method, accuracy improves by 2.06% and 1.28%, recall increases by 1.34% and 1.38%, and F1-Score rises by 1.80% and 1.21%, respectively. These advantages originate from MDCAM, which adaptively weights cross-domain features, enhances inter-domain interactions, suppresses redundancy, and exploits complementary patterns while maintaining parameter efficiency, validating the superiority and effectiveness of the proposed model.

Table 5 Comparative experimental results with other models

Model	Acc /%	R /%	F1 /%	Parameters /M	FLOPs /G
VGG16	87.30	88.16	87.81	134.27	3.72
VIT	88.58	87.42	88.09	85.67	3.42
Han et al.	90.49	88.45	88.23	44.66	2.13
Pan et al.	93.07	90.87	90.45	129.02	3.65
Wang et al.	93.85	90.83	91.04	138.56	3.93
Ours	**95.13**	**92.21**	**92.25**	85.40	2.84

Additionally, comparative validation was performed on echo signals under different Signal-to-Reverberation Ratios (SRR). The accuracy comparison of various models is shown in Fig 11. The proposed method maintains accuracy above 94% at 0-10dB SRR, while Han's model shows significant degradation. Pan's and Wang's models sustain accuracy above 92%, with the proposed method outperforming them by approximately 2%. Under low SRR conditions (0-10dB), the proposed model still achieves over 90% accuracy, surpassing all other comparative models. These results fully demonstrate that the proposed method, leveraging the MDCAM multi-domain cross-attention mechanism, effectively captures complex dependencies among multi-domain features, generates more discriminative joint representations, significantly enhances underwater target recognition accuracy, and exhibits stronger robustness in low-SRR environments.



Unedited

Fig. 11 Comparison of model accuracy under different signal-to-interference ratios

4 Conclusion

To fully exploit the multi-domain information of active sonar signals and improve the accuracy of underwater target recognition and classification, this paper proposes an attention mechanism-based multi-domain feature fusion recognition method for active sonar targets. The method achieves cross-domain information interaction of time-domain envelope features, time-azimuth features, and time-frequency features by constructing a multi-domain cross-attention mechanism. The method preprocesses active sonar echo signals and establishes a multi-domain feature extraction network, which includes a 1-DCNN-LSTM series network and a 2-DCNN network, both incorporating channel attention mechanisms to extract deep features from different domains. By combining feature concatenation and attention mechanisms for intra-domain and cross-domain feature fusion, the target classification accuracy is improved. Experimental results show that compared with single-domain methods, the network using an attention mechanism for multi-domain feature fusion enhances cross-domain information interaction, significantly improves feature representation capability, achieves notably higher recognition accuracy than other methods, and maintains stable generalization capability in low signal-to-interference ratio scenarios.

Acknowledgments

This work was supported by the Joint National Natural Science Foundation of China (No. U22A2044) and the Key Laboratory Fund from Underwater Acoustic Countermeasure Technology (No. JCKY 2024207CH01).

Author contributions

Tongjing Sun conducted formal analysis, carried out investigation, acquired resources, reviewed and edited the manuscript, acquired funding, and provided supervision. Haoran Xu designed the research concept, processed the data, carried out investigation, developed the methodology, and drafted the manuscript. Shishuo Ren carried out investigation, assisted in methodology development, and conducted validation. Denghui Zhang processed the data, conducted formal analysis, and reviewed and edited the manuscript.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Arrabito GR, Cooke BE, McFadden SM, 2005. Recommendations for enhancing the role of the auditory modality for processing sonar data. Appl Acoust, 66(8):986-1005. https://doi.org/10.1016/j.apacoust.2004.11.010

Choo Y, Lee K, Hong W, et al., 2024. Active underwater target detection using a shallow neural network with spectrogram-based temporal variation features. IEEE J Oceanic Eng, 49(1):279-293. https://doi.org/10.1109/joe.2022.3164513

Domingos LCF, Santos PE, Skelton PSM, et al., 2022. A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance. Sensors, 22(6):2181. https://doi.org/10.3390/s22062181

Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. https://doi.org/10.48550/arXiv.2010.11929

Fang SL, Du SP, Luo XW, et al., 2019. Feature analysis and recognition technology of underwater acoustic targets. Bull Chin Acad Sci, 34(3):297-305 (in Chinese). https://doi.org/10.16418/j.issn.1000-3045.2019.03.007

Gao F, Jin XP, Zhou XW, et al., 2025. MSFMamba: Multiscale feature fusion state space model for multisource remote sensing image classification. IEEE Trans Geosci Remote Sens, 63:5504116. https://doi.org/10.1109/TGRS.2025.3535622

Han XC, Ren CX, Wang LM, et al., 2022. Underwater acoustic target recognition method based on a joint neural network. PLoS ONE, 17(4):e0266425. https://doi.org/10.1371/journal.pone.0266425

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/cvpr.2016.90

He L, Liu SY, An R, et al., 2023. An end-to-end framework based on vision-language fusion for remote sensing cross-modal text-image retrieval. Mathematics, 11(10):2279. https://doi.org/10.3390/math11102279

Hong F, Liu CW, Guo L, et al., 2021. Underwater acoustic target recognition with a residual network and the optimized feature extraction method. Appl Sci, 11(4):1442. https://doi.org/10.3390/app11041442

Hu G, Wang KJ, Peng Y, et al., 2018. Deep learning methods for underwater target feature extraction and recognition. Comput Intell Neurosci, 2018:1214301. https://doi.org/10.1155/2018/1214301

Hu G, Wang KJ, Liu LL, 2021. Underwater acoustic target recognition based on depthwise separable convolution neural networks. Sensors, 21(4):1429. https://doi.org/10.3390/s21041429

Hu J, Shen L, Sun G, 2018. Squeeze-and-excitation networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7132-7141. https://doi.org/10.1109/CVPR.2018.00745

Huang HN, Li Y, 2019. Underwater acoustic detection: Current status and future trends. Bull Chin Acad Sci, 34(3):264-271 (in Chinese). https://doi.org/10.16418/j.issn.1000-3045.2019.03.003

Kamal S, Chandran CS, Supriya MH, 2021. Passive sonar automated target classifier for shallow waters using end-to-end learnable deep convolutional LSTMs. Eng Sci Technol Int J, 24(4):860-871.

Unedited

https://doi.org/10.1016/j.jestch.2021.01.014

Khan A, Fouda MM, Do DT, et al., 2024. Underwater target detection using deep learning: methodologies, challenges, applications, and future evolution. IEEE Access, 12:12618-12635. https://doi.org/10.1109/ACCESS.2024.3353688

Lee S, Seo I, Seok J, et al., 2020. Active sonar target classification with power-normalized cepstral coefficients and convolutional neural network. Appl Sci, 10(23):8450. https://doi.org/10.3390/app10238450

Li HC, Lu YT, Zhu HD, 2024. Multi-modal sentiment analysis based on image and text fusion based on cross-attention mechanism. Electronics, 13(11):2069. https://doi.org/10.3390/electronics13112069

Liu DL, Zhao XC, Cao WJ, et al., 2020. Design and performance evaluation of a deep neural network for spectrum recognition of underwater targets. Comput Intell Neurosci, 2020:8848507. https://doi.org/10.1155/2020/8848507

Pan XY, Sun J, Feng TH, et al., 2025. Underwater target recognition based on adaptive multi-feature fusion network. Multimed Tools Appl, 84(10):7297-7317. https://doi.org/10.1007/s11042-024-19178-9

Shadlou Jahromi M, Bagheri V, Rostami H, et al., 2019. Feature extraction in fractional Fourier domain for classification of passive sonar signals. J Signal Process Syst, 91(5):511-520. https://doi.org/10.1007/s11265-018-1347-x

Shin FB, Kil DH, Wayland RF, 1997. Active impulsive echo discrimination in shallow water by mapping target physics-derived features to classifiers. IEEE J Oceanic Eng, 22(1):66-80. https://doi.org/10.1109/48.557541

Simonyan K, Zisserman A, 2015. Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv.1409.1556

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010. https://doi.org/10.48550/arXiv.1706.03762

Wang QC, Du SP, Zhang W, et al., 2024. Active sonar target recognition method based on multi-domain transformations and attention-based fusion network. IET Radar Sonar Nav, 18(10):1814-1828. https://doi.org/10.1049/rsn2.12618

Yang HH, Gan AQ, Chen HL, et al., 2016. Underwater acoustic target recognition using SVM ensemble via weighted sample and feature selection. 13th Int Bhurban Conf on Applied Sciences and Technology, p.522-527. https://doi.org/10.1109/ibcast.2016.7429928

Young VW, Hines PC, 2007. Perception-based automatic classification of impulsive-source active sonar echoes. J Acoust Soc Am, 122(3):1502-1517. https://doi.org/10.1121/1.2767001

Zhang W, Wu YQ, Wang DZ, et al., 2018. Underwater target feature extraction and classification based on gammatone filter and machine learning. Int Conf on Wavelet Analysis and Pattern Recognition, p.42-47. https://doi.org/10.1109/icwapr.2018.8521356

Unedited